THE CORRELATION BETWEEN A COMPONENT, AND BETWEEN THE SUM OF TWO OR MORE COMPONENTS, AND THE SUM OF THE REMAINING COMPONENTS OF A VARIABLE.

By J. Arthur Harris, *Carnegie Institution of Washington, Station for Experimental Evolution, Cold Spring Harbor, N. Y.*

Many quantitatively measured variables with which the statistician has to deal in both social and biological sciences are really composite in character. Thus the death rates of a series of districts are made up of mortalities due to a number of causes. The total milk production of a cow is the sum of the productions of individual lactation periods which may differ greatly in quantity. The total length of an organism is the sum of the lengths of its component parts. The annual egg production of a fowl in an egg laying competition is the sum of the records for individual months.

Thus x_1, x_2, x_3 . . . x_n are the components of the variable X, where $X=\Sigma(x)$.

It is customary, and in many instances quite proper, to deal with all such constituent elements as quite independent variables. When only one correlation involving two individual components is to be determined, a table may be formed in the usual manner. Cases may, however, arise in statistical analysis in which it is desirable to determine the correlation between the magnitude of any individual constituent element, x, of the variable X and the sum of the remaining elements $(X-x)$. Indeed, all possible measures of this kind may be needed, *i. e.*, $r_{x_1(X-x_1)}$, $r_{x_2(X-x_2)}$, $r_{x_3(X-x_3)}$, . . . $r_{x_n(X-x_n)}$.

If X be large and the values of x_1, x_2, x_3 . . . x_n variable, the arithmetical routine may be troublesome, in fact practically prohibitive.

The correlation between the sum of two components and the sum of the remaining components

$$r_{(x_1+x_2)(X-x_1-x_2)},\ r_{(x_1+x_3)(X-x_1-x_3)}\ \cdot\ \cdot\ \cdot\ r_{(x_1+x_n)(X-x_1-x_n)},$$
$$\cdot\ \cdot\ \cdot\ r_{(x_2+x_3)(X-x_2-x_3)},\ \cdot\ \cdot\ \cdot\ r_{(x_{n-1}+x_n)(X-x_{n-1}-x_n)},$$

or between the sum of three or more components and the sum of the remaining components, *e. g.*,

$$r_{(x_1+x_2+x_3)(X-x_1-x_2-x_3)},\ r_{(x_2+x_3+x_4)(X-x_2-x_3-x_4)},\ \cdot\ \cdot\ \cdot$$
$$r_{(x_{n-2}+x_{n-1}+x_n)(X-x_{n-2}-x_{n-1}-x_n)},$$

or any of the other possible permutations which increase rapidly with the number of components, may be quite as important, at least, as the correlation between any individual component and the sum of the remaining components.

The determination of all such relationships is relatively simple providing the correlations between the components are known or their product moments have been determined.

Since these methods may not suggest themselves to those not familiar with the many possible modifications of the correlation formulae it seems worth while to put the equations on record.

The first requisite is the means and standard deviations of the sum of two or more components and of the difference between the variable and the sum of two or more components. The principle is well known.*

The value of σ_X is known and constant throughout. The variable values of σ_{x_1}, σ_{x_2}, σ_{x_3} . . . σ_{x_n} and r_{x_1X}, r_{x_2X}, r_{x_3X} . . . r_{x_nX} have also been determined. The standard deviation for the variable less any component, say the pth, is

$$\sigma^2_{(X-x_p)} = \sigma^2_X + \sigma^2_{x_p} - 2r_{x_pX}\sigma_X\sigma_{x_p}$$

For the sum of any number of components the standard deviation is given by

$$\sigma^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 + \ \cdot\ \cdot\ \cdot\ + \sigma_{x_n}^2 + 2r_{x_1x_2}\sigma_{x_1}\sigma_{x_2} + 2r_{x_1x_3}\sigma_{x_1}\sigma_{x_3}$$
$$+\ \cdot\ \cdot\ \cdot\ 2r_{x_2x_3}\sigma_{x_2}\sigma_{x_3} +\ \cdot\ \cdot\ \cdot$$

In actual practise it is often most convenient to work from the original summations.† Thus the moments $\Sigma(x_1)$, $\Sigma(x_1^2)$,

* See for example, R. Pearl, Biometrika, Vol. VI, pp. 437–438, 1909, and G. U. Yule, Introduction to the Theory of Statistics, p. 208, 1911.

† Harris, J. Arthur. The arithmetic of the product moment method of calculating the coefficient of correlation. *American Naturalist*, Vol. XLIV, pp. 693–699, 1910.

$\Sigma(x_2)$, $\Sigma(x_2^2)$, $\Sigma(x_3)$, $\Sigma(x_3^2)$ . . . $\Sigma(x_n)$, $\Sigma(x_n^2)$ and $\Sigma(X)$, $\Sigma(X^2)$ with the product moments $\Sigma(x_1X)$, $\Sigma(x_2X)$, $\Sigma(x_3X)$ . . . $\Sigma(x_nX)$ lead directly to the desired results.

In the following section I give the equations in terms of the original moments or moment coefficients.

The means and standard deviations for the individual components, say for example the pth component, are

$$\bar{x}_p=\Sigma(x_p)/N,\ \sigma^2_{x_p}=\Sigma(x_p^2)/N-[\Sigma(x_p)/N]^2$$

The means of the sums of the remaining components are quite obviously

$$\overline{(X-x_1)}=[\Sigma(X)-\Sigma(x_1)]/N,\text{ etc.,}$$

while the standard deviations are given by

$$\sigma^2{}_{(X-x_1)}=\{\Sigma(X^2)-2\Sigma(x_1X)+\Sigma(x_1^2)\}/N-(\overline{X-x_1})^2,$$
$$\sigma^2{}_{(X-x_2)}=\{\Sigma(X^2)-2\Sigma(x_2X)+\Sigma(x_2^2)\}/N-(\overline{X-x_2})^2,\text{ etc.}$$

The mean for the sum of two components is

$$\overline{(x_p+x_q)}=[\Sigma(x_p)+\Sigma(x_q)]/N,$$

and so on for any number of components.

For the sum of the remaining $(n-2)$, $(n-3)$ components the means are

$$\overline{(X-x_p-x_q)}=[\Sigma(X)-\Sigma(x_p)-\Sigma(x_q)]/N,$$

and similarly for reductions due to the removal of 3, 4 or more components.

For the sum of two components, (x_p+x_q)

$$\sigma^2=[\Sigma(x_p^2)+2\Sigma(x_px_q)+\Sigma(x_q^2)]/N-(\overline{x_p+x_q})^2$$

For three components, $(x_m+x_p+x_q)$

$$\sigma^2=[\Sigma(x_m^2)+\Sigma(x_p^2)+\Sigma(x_q^2)+2\Sigma(x_mx_p)+2\Sigma(x_mx_q)$$
$$+2\Sigma(x_px_q)]/N-(\overline{x_m+x_p+x_q})^2$$

For four components, $(x_h+x_m+x_p+x_q)$

$$\sigma^2=[\Sigma(x_h^2)+\ \ .\ .\ .\ \ +\Sigma(x_q^2)+2\Sigma(x_hx_m)+\ \ .\ .\ .$$
$$+2\Sigma(x_px_q)]/N-(\overline{x_h+\ \ .\ .\ .\ \ +x_q})^2,$$

and so on for higher numbers of components.

The values of the means of the $(n-2)$, $(n-3)$, $(n-4)$ remaining components have been indicated.

The standard deviation of the value which remains after deducting the values of two components x_p and x_q is

$$\sigma^2=[\Sigma(X^2)+\Sigma(x_p{}^2)+\Sigma(x_q{}^2)-2\Sigma(x_pX)-2\Sigma(x_qX)+2\Sigma(x_px_q)]/N \\ -\overline{(X-x_p-x_q)}^2$$

The value of $\sigma_{(X-x_m-x_p-x_q)}$ is given by

$$\sigma^2=[\Sigma(X^2)+\Sigma(x_m{}^2)+\Sigma(x_p{}^2)+\Sigma(x_q{}^2)-2\Sigma(x_mX)- \\ 2\Sigma(x_pX)-2\Sigma(x_qX)+2\Sigma(x_mx_p)+2\Sigma(x_mx_q)+ \\ 2\Sigma(x_px_q)]/N-\overline{(X-x_m-x_p-x_q)}^2$$

The process for the value remaining after the renewal of four components is

$$\sigma^2=[\Sigma(X^2)+\Sigma(x_h{}^2)+\quad.\quad.\quad.\quad+\Sigma(x_q{}^2)-2\Sigma(x_hX)- \\ .\quad.\quad.\quad-2\Sigma(x_qX)+2\Sigma(x_hx_m)+\quad.\quad.\quad. \\ +2\Sigma(x_px_q)]/N-\overline{(X-x_h-\quad.\quad.\quad.\quad-x_q)}^2$$

I now turn to the correlations.

Consider first of all the simplest of the two problems, the determination of the correlation between a component and the sum of the remaining components of the variable.

The correlations between the variable and its constituent elements r_{x_1X}, r_{x_2X}, r_{x_3X}, . . . r_{x_nX} are often wanted for themselves, and in any instance are relatively easily determined. The regressions of X on its constituent elements, say on the pth component, is

$$r_{x_pX}\frac{\sigma_X}{\sigma_{x_p}}$$

Obviously the regression slope of $(X-x_p)$ on x_p is

$$r_{x_p(X-x_p)}\frac{\sigma_{(X-x_p)}}{\sigma_{x_p}}=r_{x_pX}\frac{\sigma_X}{\sigma_{x_p}}-1,$$

or in terms of correlation

$$r_{x_p(X-x_p)}=r_{x_pX}\frac{\sigma_X}{\sigma_{(X-x_p)}}-\frac{\sigma_{x_p}}{\sigma_{(X-x_p)}}$$

I now turn to the formulae necessary for the determination of the correlation between the sum of any two components, say x_p and x_q, and the sum of the remaining components,

$(X-x_p-x_q)$, of any three components, say x_m, x_p and x_q, and the sum of the remaining components $(X-x_m-x_p-x_q)$, and between the sum of any four components, say x_h, x_m, x_p and x_q and the sum of the remaining components, $(X-x_h-x_m-x_p-x_q)$.

The product moments for the sum of two variables (x_p+x_q) and the sum of the remaining components $(X-x_p-x_q)$ is

$$\Sigma(x_pX)+\Sigma(x_qX)-2\Sigma(x_px_q)-\Sigma(x_p^2)-\Sigma(x_q^2)$$

For the sum of three components, $(x_m+x_p+x_q)$, and the sum of the remaining components, $(X-x_m-x_p-x_q)$, the product moment is

$$\Sigma(x_mX)+\Sigma(x_pX)+\Sigma(x_qX)-2\Sigma(x_mx_p)-2\Sigma(x_mx_q)-2\Sigma(x_px_q)$$
$$-\Sigma(x_m^2)-\Sigma(x_p^2)-\Sigma(x_q^2)$$

The product moment for four components $(x_h+x_m+x_p+x_q)$ and the sum of the remaining components is of course

$$\Sigma[(x_h+x_m+x_p+x_q)(X-x_h-x_m-x_p-x_q)]$$

which is easily thrown into the convenient form illustrated above for two and three components.

In certain cases the coefficient of correlation between a single component, say x_p, and the sum of the components remaining after the deduction of the *p*th and one or more other components is desired. Or, more generally, the coefficient of correlation between the sum of *n* components and the sum of the components remaining after the deduction of $n+1$, $n+2$, $n+3$ . . . components may be required.

The moments from which the means and standard deviations of the various components, sums of components and differences may be computed, have already been indicated. Thus the product moments only are required.

For the correlation between a single component, x_p, and the variable less two components, x_p and x_q, the product moment is

$$\Sigma(x_pX)-\Sigma(x_p^2)-\Sigma(x_px_q)$$

For the correlation between a component, x_p, and the variable less three components, $(X-x_p-x_q-x_m)$, the product moment is

$$\Sigma(x_pX)-\Sigma(x_p^2)-\Sigma(x_px_q)-\Sigma(x_px_m)$$

The formula for the product moment underlying the correlation between a component and the variable less the sum of a larger number of components is obvious.

For the sum of two components, (x_p+x_q), and the value of the variable less three components, $(X-x_p-x_q-x_m)$, the product moment is

$$\Sigma(x_pX)+\Sigma(x_qX)-\Sigma(x_p^2)-\Sigma(x_q^2)-2\Sigma(x_px_q)$$
$$-\Sigma(x_px_m)-\Sigma(x_qx_m)$$

Correlations beyond these for the sum of three components, $(x_p+x_q+x_m)$, and the value of the variable less that of four components, $(X-x_p-x_q-x_m-x_h)$, will be required very rarely indeed. The product moment for such a relationship is

$$\Sigma(x_pX)+\Sigma(x_qX)+\Sigma(x_mX)-\Sigma(x_p^2)-\Sigma(x_q^2)-\Sigma(x_m^2)-2\Sigma(x_px_q)$$
$$-2\Sigma(x_px_m)-2\Sigma(x_qx_m)-\Sigma(x_px_h)-\Sigma(x_qx_h)-\Sigma(x_mx_h)$$

Thus the formulae for the determination of all of the fundamental constants required in the calculation of the coefficient of correlation between the sum of from two to four components and the sum of the remaining components have been deduced.

Some of these results are known, others are such as anyone familiar with elementary statistical theory might write for himself. My object here has not been to express the relationships in the most elegant form but in that most convenient for practical work. The computer will note the great advantages of determining the moments and product moments for the individual components once for all. It is often just such simplification of method that greatly decreases the labor involved in the routine of the computing room.